SRUTHI SATYAVARAPU

M.S. Computer Science — Applied AI Engineer | AI/ML Engineer | Software Engineer

mailto:sruthiraosatyavarapu@gmail.com | tel: +1 510-365-2100 | San Jose, CA
https://linkedin.com/in/sruthi-satyavarapu | https://github.com/sruthisDev | https://sruthirao.com

ABOUT

AI/ML Engineer and Software Engineer with 7+ years of experience designing and deploying scalable, cloud-native intelligent systems across healthcare, finance, and e-commerce. Combines a strong backend engineering foundation - APIs, microservices, cloud deployment on AWS with hands on ML expertise in LLMs, RAG, NLP, and end-to-end model evaluation. Currently pursuing M.S in Computer Science, while actively contributing to production-grade AI systems research.

EDUCATION

M.S. Computer Science
University of the Pacific, Stockton
GPA 4.0 | Aug 2024 – Present

M.S. Technology (Medical Software)
Manipal Institute of Technology, India
GPA 3.7 | Aug 2014 – May 2016

TECHNICAL SKILLS

- **Programming:** Python, C++, JavaScript, TypeScript, PHP
- **AI/ML:** Machine Learning, NLP, LLMs, RAG, Model Evaluation, Model Fine-Tuning, Hugging Face, Scikit-learn, LangChain, PyTorch, RAGAS
- **Backend/Data Engineering:** FastAPI, Flask, ETL Pipelines, Microservices, REST APIs, Airflow, Spring Boot, Data Preprocessing.
- **Frontend:** React.js, HTML/CSS
- **Databases:** MySQL, MongoDB, ChromaDB, SQL
- **Cloud/DevOps:** AWS(EC2, S3, SageMaker, Lambda) , Docker, CI/CD, Kubernetes
- **Security:** OWASP, OAuth2, JWT Auth, RBAC
- **Tools:** Postman, Jupyter, Git, Tableau

RESEARCH EXPERIENCE

Research Assistant – Retrieval-Augmented Generation Systems (March 2025 – Present) - University of the Pacific

- Designed and implemented a modular RAG architecture addressing the gap between benchmarking frameworks like FlashRAG and deployment frameworks (LangChain, LlamaIndex), enabling configuration-driven experimentation without architectural changes.
- Built end-to-end system using ReactJS, FastAPI, and ChromaDB incorporating query rewriting, re-ranking, session-aware interaction, and RBAC-based access control for multi-user environments. Integrated persistent storage and latency benchmarking infrastructure for reproducible cross-domain RAG evaluation.
- Evaluated retrieval performance across biomedical and Sanskrit philosophical corpora using Recall@K, MRR, and RAGAS, demonstrating high domain sensitivity to component selection.
- *Paper under review - IEEE BigData Service: 'A Modular Architecture for Domain-Adaptive Retrieval-Augmented Generation Systems'*

Research Assistant – Machine Learning for IoT Systems (Jun 2025 – Dec 2025) - University of the Pacific

- Developed ML models using PyTorch and Scikit-learn to analyze wearable sensor data streams for IoT-based health monitoring, improving anomaly detection accuracy by 28%.
- Engineered time-series feature extraction pipelines for environmental and motion-based datasets, enhancing predictive model performance across multiple sensor modalities.
- Built Flask-based RESTful services integrated with React dashboards to deliver real-time sensor insights and model outputs to end users.

INDUSTRY EXPERIENCE

AI/ML Intern – PNC, USA (Jan 2024 – July 2024)

- Built end-to-end ML pipelines using AWS SageMaker, and Airflow for financial risk systems, improving processing efficiency by 30%.
- Developed NLP-based fraud detection models leveraging LLMs and ensemble ML techniques, reducing detection latency by 15% and improving real-time risk scoring accuracy.
- Designed and deployed scalable ML services via Java Spring Boot microservices, Docker, and Kubernetes across cloud environments, ensuring high availability and seamless enterprise integration.

Software Developer – Tata Consultancy Services, Hyderabad (Jul 2021 – April 2023)

- Developed and deployed ML models for predictive analytics on the Mars Petcare platform, improving forecasting accuracy by 25%.
- Implemented secure backend systems following OWASP guidelines for enterprise applications, improving performance by 20% and reducing vulnerabilities by 35%.
- Deployed and maintained scalable cloud solutions on AWS (EC2, S3, Lambda) supporting high availability.

Software Engineer – Hyper Interact, Foray, PurpleTalk, Excellera (2016 – 2021, India)

- Delivered web and mobile solutions across bioinformatics, e-commerce, and digital services domains for multiple clients, receiving performance awards for productivity improvements up to 25%.
- Built backend systems and delivered scalable solutions across PHP, MySQL, and MongoDB stacks, establishing core practices in database design and workflow automation.

PROJECTS

AI Companion for Seniors [Ongoing]
AWS Bedrock, MongoDB, FastAPI, React, Python
Voice-enabled AI health companion using AWS Bedrock and RAG, with medication tracking, anomaly detection, and real-time caregiver dashboard.

Strawberry Ripeness Detection
Python, Scikit-learn
Ensemble model (KNN + SVM) achieving 92% accuracy for automated agricultural quality assessment.

Weather–Music Trends Analysis
Python, Tableau
Performed EDA across 100+ cities to uncover weather–music correlations and genre trends.

Secure Hospital Management System
React, Node.js, MySQL, OAuth2, OWASP
Role-based hospital management system for patients and doctors with OAuth2 authentication and secure API design following OWASP guidelines.